Tomographic inversion using ℓ_1-norm regularization of wavelet coefficients

Ignace Loris[1,2], Guust Nolet[3], Ingrid Daubechies[1], F. A. Dahlen[3]

[1]Program in Applied and Computational Mathematics,
Princeton University, Princeton, New Jersey, USA
[2]Dienst Theoretische Natuurkunde, Vrije Universiteit Brussel,
Brussels, Belgium, E-mail: igloris@vub.ac.be
[3]Department of Geosciences, Princeton University, Princeton, New Jersey, USA

February 2, 2008

Abstract

We propose the use of ℓ_1 regularization in a wavelet basis for the solution of linearized seismic tomography problems $\mathbf{Am} = \mathbf{d}$, allowing for the possibility of sharp discontinuities superimposed on a smoothly varying background. An iterative method is used to find a sparse solution $\mathbf{m}$ that contains no more fine-scale structure than is necessary to fit the data $\mathbf{d}$ to within its assigned errors. keywords: inverse problem, one-norm, sparsity, tomography, wavelets

1 Introduction

Like most geophysical inverse problems, the linearized problem $\mathbf{Am} = \mathbf{d}$ in seismic tomography is underdetermined, or at best offers a mix of overdetermined and underdetermined parameters. It has therefore long been recognized that it is important to suppress artifacts that could be falsely interpreted as 'structure' in the earth's interior. Not surprisingly, strategies that yield the smoothest solution $\mathbf{m}$ have been dominant in most global or regional tomographic applications; these strategies include seeking global models represented as a low-degree spherical harmonic expansion [Dziewonski et al.(1975), Dziewonski & Woodhouse(1987), Masters et al.(1996)] as well as regularization via minimization of the gradient ($\nabla\mathbf{m}$) or second derivative ($\nabla^2\mathbf{m}$) norm of a dense local parametrization [Nolet(1987), Constable et al.(1987), Spakman & Nolet(1988), VanDecar & Snieder(1994), Trampert & Snieder(1996)].

Smooth solutions, however, while not introducing small-scale artifacts, produce a distorted image of the earth through the strong averaging over large areas, thereby making small-scale detail difficult to see, or even hiding it. Sharp discontinuities are blurred into gradual transitions. For example, the inability of global, spherical-harmonic, tomographic models to yield as clear an image of upper-mantle subduction zones as produced by more localized studies has long been held against them. [Deal et al.(1999)] and [Deal & Nolet(1999)] optimize images of upper-mantle slabs to fit physical models of heat diffusion, in an effort to suppress small-scale imaging artifacts while retaining sharp boundaries. [Portniaguine & Zhdanov(1999)] use a conjugate-gradient method to seek the smallest possible anomalous domain by minimizing a norm based on a renormalized gradient $\nabla\mathbf{m}/(\nabla\mathbf{m}\cdot\nabla\mathbf{m}+\gamma^2)^{\frac{1}{2}}$, where γ is a small constant. Like all methods that deviate from a least-squares type of solution, both these methods are nonlinear and pose their own problems of practical implementation.

The notion that we should seek the 'simplest' model $\mathbf{m}$ that fits a measured set of data $\mathbf{d}$ to within the assigned errors is intuitively equivalent to the notion that the model should be describable with a

small number of parameters. But, clearly, restricting the model to a few low-degree spherical-harmonic or Fourier coefficients, or a few large-scale blocks or tetrahedra, does not necessarily lead to a geophysically plausible solution. In this paper we investigate whether a multiscale representation based upon wavelets [Daubechies(1992)] has enough flexibility to represent the class of models we seek. We propose an ℓ_1-norm regularization method which yields a model $\mathbf{m}$ that has a strong tendency to be *sparse* in a wavelet basis, meaning that it can be faithfully represented by a relatively small number of nonzero wavelet coefficients. This allows for models that vary smoothly in regions of limited coverage without sacrificing any sharp or small-scale features in well-covered regions that are required to fit the data. Our approach is different from an approach briefly suggested by [de Hoop & van der Hilst(2005)], in which the mapping between data and model is decomposed in curvelets: here we are concerned with applying the principle of parsimony to the solution of the inverse problem, without any special preference for singling out linear features, for which curvelets are probably better adapted than wavelets.

In Section 2 we give a short description of the mathematical method, and in Section 3 we consider a geophysically motivated, toy 2D application, in which the synthetic data are a small set of regional, fundamental-mode, Rayleigh-wave dispersion measurements expressed as wavenumber perturbations $\delta k(\nu)$ at various frequencies ν. To enable us to concentrate on the mathematical rather than the geophysical aspects of the inverse problem, we assume that the fractional shear-velocity perturbations $\delta\ln\beta = \delta\beta/\beta$ within the region are depth-independent. Finite-frequency interpretation of the surface-wave dispersion data [Zhou et al.(2004)] then yields a 2D linearized inverse problem of the form $\mathbf{Am} = \mathbf{d}$. We compare wavelet-basis models $\mathbf{m}$ obtained using our proposed ℓ_1-norm regularization with models obtained using more conventional ℓ_2 regularization, both with and without wavelets, and show that the former are sparser and have fewer small-scale artifacts.

2   Mathematical principles

In any realistic tomographic problem, the linear system $\mathbf{Am} = \mathbf{d}$ is not invertible: even when the number of data exceeds the number of unknowns, the least-squares matrix $\mathbf{A}^T\mathbf{A}$ is (numerically) singular. Additional conditions always have to be imposed. The proposed regularization method is based on the fundamental assumption that the model $\mathbf{m}$ is sparse in a wavelet basis [Daubechies(1992)]. We believe that this is an appropriate inversion philosophy for finding a smoothly varying model while still allowing for whatever sharp or small-scale features are required to fit the data $\mathbf{d}$. An important feature of the method is that the location of the small-scale features does not have to be specified beforehand.

A wavelet decomposition is a special kind of basis transformation that can be computed efficiently (the number of operations is proportional to the number of components in the input). At each step the algorithm strips off detail belonging to the finest scale present —this detail is encoded in wavelet coefficients, broadly corresponding to local differences— and calculates a coarse version —encoded in scaling coefficients, broadly corresponding to local averages— that is only half the size of the original in 1D and only one quarter the size in 2D. This procedure is repeated on the successive coarse versions. The resulting wavelet coefficients (at the different scales) and scaling coefficients (at the final coarsest scale only) are called the wavelet decomposition of the input. By this construction each wavelet coefficient carries information belonging to a certain scale (by virtue of the decimation) and a certain position (use of *local* differences). The final few scaling coefficients represent a (very) coarse average.

The mathematical relation between the wavelet-basis expansion coefficients $\mathbf{w}$ and the model $\mathbf{m}$ is the wavelet transform $\mathbf{W}$ (a linear operator): $\mathbf{m} = \mathbf{W}^T\mathbf{w}$. By choosing the local differences and averages carefully (corresponding to a choice among many different so-called wavelet families), the inverse transformation from $\mathbf{w}$ back to $\mathbf{m}$ can be made equally efficient. In our application we will use a special kind of 2D wavelet basis that is overcomplete: it contains six different wavelets corresponding to different directions. Because of this overcompleteness, the wavelet transform $\mathbf{W}$ has a left inverse (namely $\mathbf{W}^T$): $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, but no right inverse, $\mathbf{W}\mathbf{W}^T \neq \mathbf{I}$. Appendix B contains a short overview of this particular
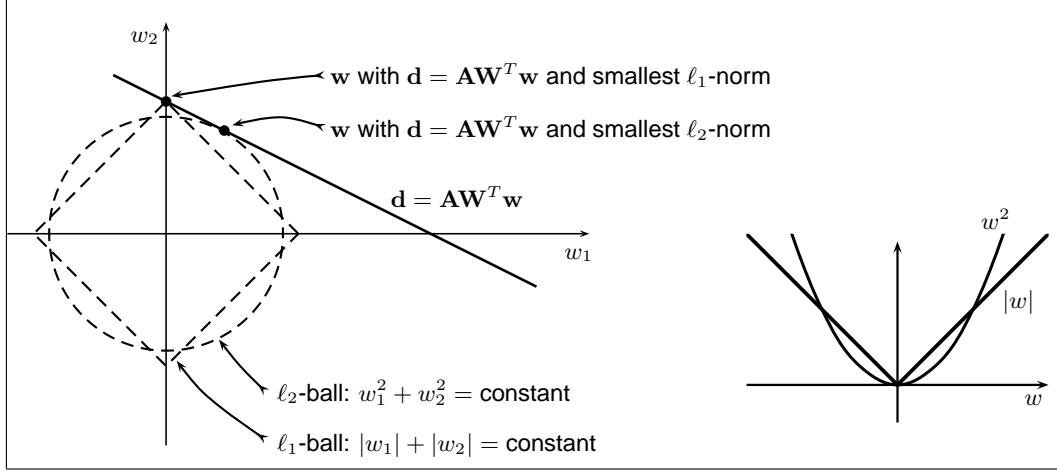
Figure 1: Sparsity, ℓ_1 minimization and ℓ_2 minimization: Left: Because the ℓ_1-ball has no bulge, the solution with smallest ℓ_1-norm is sparser than the solution with smallest ℓ_2-norm. Right: A $|w|$ penalization effects small coefficients more and large coefficients less than the (traditional) w^2 penalization.

construction. In short our wavelet and scaling coefficients $\mathbf{w}$ contain information on scale, position and direction.

For the tomographic reconstruction, we will require a sparse set of wavelet-basis coefficients: the vast majority of these represent differences and will only be present around non-smooth features. In this way we regularize the inversion by adapting ourselves to the model rather than to the operator. As a measure of sparsity we will use the ℓ_1-norm of the wavelet representation $\mathbf{w}$ of the model $\mathbf{m}$, i.e. we will look for a solution of the linear equations $\mathbf{Am} = \mathbf{d}$ that has a small $\|\mathbf{w}\|_1 = \sum_i |w_i|$. Since $|w_i| > |w_i|^2$ for small w_i and $|w_i| < |w_i|^2$ for large w_i, this type of penalization will favor a small number of large coefficients over a large number of small coefficients in the reconstruction (whereas a traditional ℓ_2 penalization might do the opposite). We are not claiming that the sparsest solution always coincides with the minimum ℓ_1-norm solution, but one can show that it often does [Donoho(2004), Candes et al.(2006)]. A schematic justification for this is given in Fig. 1.

In particular, our strategy will consist of searching for the minimizer of the functional

$$I_1(\mathbf{w}) = \|\mathbf{d} - \mathbf{Am}\|_2^2 + 2\tau \|\mathbf{w}\|_1 = \|\mathbf{d} - \mathbf{A}\mathbf{W}^T\mathbf{w}\|_2^2 + 2\tau \|\mathbf{w}\|_1, \tag{1}$$

where τ is an adjustable parameter at our disposal. Here, the first (quadratic) term corresponds to the conventional statistical measure of misfit to the data, $\chi^2 = \sum_i [d_i - (\mathbf{Am})_i]^2$, and the second ($\ell_1$-norm) term $2\tau \|\mathbf{w}\|_1$ is introduced to regularize the inversion. In writing χ^2 in this form, we have made the simplifying assumption that the noisy data $\mathbf{d}$ are uncorrelated with unit variance. More generally, the misfit portion of the functional (1) is $\chi^2 = (\mathbf{d} - \mathbf{Am})^T \boldsymbol{\Sigma}^{-1}(\mathbf{d} - \mathbf{Am})$, where $\boldsymbol{\Sigma}$ is the data covariance matrix. In the 2D toy problem considered in Section 3, we invert synthetic data $\mathbf{d}$ having a constant (but non-unit) variance, $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$.

The minimizer of the functional (1) can be found by iteration [Daubechies et al.(2004)]: starting with the present approximation $\mathbf{w}^{(n)}$ one constructs an nth-iterate surrogate functional

$$I_1^{(n)}(\mathbf{w}) = I_1(\mathbf{w}) - \|\mathbf{A}\mathbf{W}^T(\mathbf{w} - \mathbf{w}^{(n)})\|_2^2 + \|\mathbf{w} - \mathbf{w}^{(n)}\|_2^2 \tag{2}$$

that has the same value and the same derivative at the point $\mathbf{w} = \mathbf{w}^{(n)}$ as the original functional (see

Fig. 2). This surrogate functional can be rewritten as

$$I_1^{(n)}(w) = \left\| \mathbf{w} - \left(\mathbf{WA}^T\mathbf{d} + (\mathbf{I} - \mathbf{WA}^T\mathbf{AW}^T)\mathbf{w}^{(n)} \right) \right\|_2^2 + 2\tau\|\mathbf{w}\|_1 + c^{(n)}, \tag{3}$$

where $c^{(n)}$ is independent of $\mathbf{w}$. This functional has a much simpler form than the original $I_1(\mathbf{w})$ because there is no operator $\mathbf{AW}^T$ mixing different components of $\mathbf{w}$. The next approximation $\mathbf{w}^{(n+1)}$ is defined by the minimizer of this new functional. By calculating the derivative of expression (3) with respect to a specific wavelet or scaling coefficient w_i, one finds the following set of component-by-component equations:

$$w_i - \left(\mathbf{WA}^T\mathbf{d} + (\mathbf{I} - \mathbf{WA}^T\mathbf{AW}^T)\mathbf{w}^{(n)} \right)_i + \tau\,\mathrm{sign}(w_i) = 0, \tag{4}$$

valid whenever $w_i \neq 0$. These equations are solved by distinguishing the two cases $w_i > 0$ and $w_i < 0$; the solution —corresponding to the minimizer of the surrogate functional $I_1^{(n)}(w)$, and denoted by $\mathbf{w}^{(n+1)}$— is then found to equal

$$\mathbf{w}^{(n+1)} = \mathcal{S}_\tau\left[\mathbf{WA}^T\mathbf{d} + (\mathbf{I} - \mathbf{WA}^T\mathbf{AW}^T)\mathbf{w}^{(n)} \right], \tag{5}$$

where $\mathcal{S}_\tau$ is the so-called soft-thresholding operation, i.e.

$$\mathcal{S}_\tau(w) = \begin{cases} w - \tau & w \geq \tau \\ 0 & |w| \leq \tau \\ w + \tau & w \leq -\tau, \end{cases} \tag{6}$$

performed on each wavelet or scaling coefficient w_i individually. The starting point of the iteration procedure is arbitrary, e.g. $\mathbf{w}^{(0)} = \mathbf{0}$. Because of the component-wise character of the tresholding, it is straightforward to use different thresholds τ_i for different components w_i if desired, and in fact we shall use different thresholds τ_w and τ_s for the wavelet and scaling coefficients in our application. A schematic representation of the idea behind the iteration (5) is given in Fig. 2. We realize that this iteration converges slowly for ill-conditioned matrices, but we use it here because it is proven to converge to the solution [Daubechies et al.(2004)].

An improvement in convergence can be gained by rescaling the operator $\mathbf{A}$ (and rescaling the data $\mathbf{d}$ at the same time) in such a way that the largest eigenvalue of $\alpha^2\mathbf{A}^T\mathbf{A}$ is close to (but smaller than) unity. The iteration corresponding to the minimization of this new, rescaled functional is

$$\mathbf{w}^{(n+1)} = \mathcal{S}_{\tau\alpha^2}\left[\alpha^2\mathbf{WA}^T\mathbf{d} + (\mathbf{I} - \alpha^2\mathbf{WA}^T\mathbf{AW}^T)\mathbf{w}^{(n)} \right]. \tag{7}$$

We will also make use of the following two-step procedure: from the outcome $\overline{\mathbf{m}} = \mathbf{W}^T\overline{\mathbf{w}}$ of the iteration (7), we define new, linearly shifted data $\mathbf{d}' = 2\mathbf{d} - \mathbf{A}\overline{\mathbf{m}}$ and restart the same iteration with this new data:

$$\mathbf{w}^{(n+1)} = \mathcal{S}_{\tau\alpha^2}\left[\alpha^2\mathbf{WA}^T\mathbf{d}' + (\mathbf{I} - \alpha^2\mathbf{WA}^T\mathbf{AW}^T)\mathbf{w}^{(n)} \right], \qquad \mathbf{w}^{(0)} = \overline{\mathbf{w}}. \tag{8}$$

The outcome $\overline{\overline{\mathbf{m}}} = \mathbf{W}^T\overline{\overline{\mathbf{w}}}$ of this second iteration is then the final, regularized reconstruction of the model. For the same value of the regularization parameter τ, the second step improves the data fit considerably, $\|\mathbf{d} - \mathbf{A}\overline{\overline{\mathbf{m}}}\|_2^2 < \|\mathbf{d} - \mathbf{A}\overline{\mathbf{m}}\|_2^2$; hence a given level of final data fit χ^2 will, in the two-step procedure, correspond to a higher value of τ. Because τ specifies the threshold level, a higher value will lead to more aggressive thresholding and thus faster convergence to a sparse solution.

The above method will be demonstrated in the next section and compared to a conventional ℓ_2-regularization method, in which the functional

$$I_2(\mathbf{m}) = \|\mathbf{d} - \mathbf{Am}\|_2^2 + \tau\|\mathbf{m}\|_2^2 \tag{9}$$
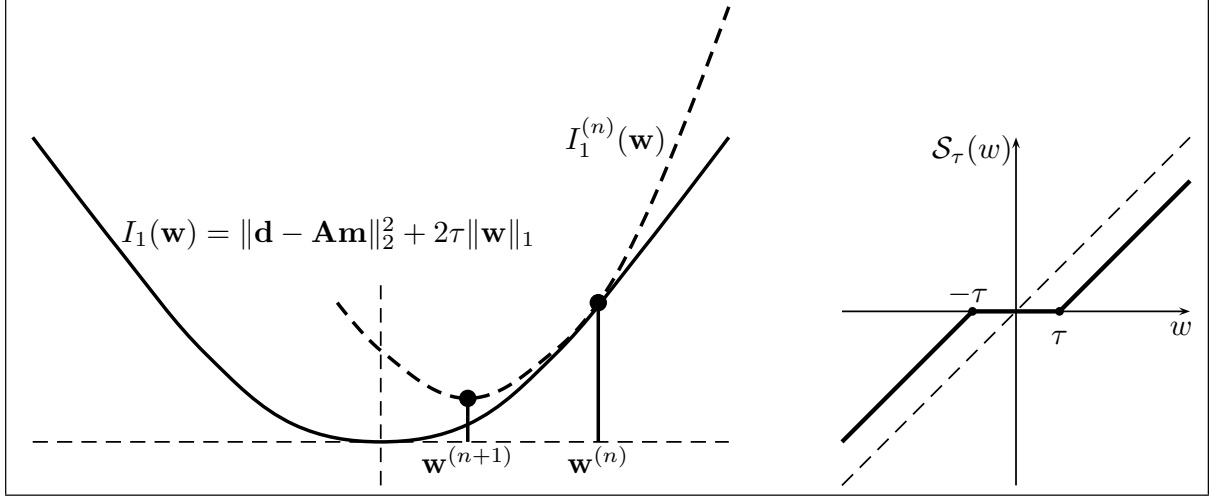
4

Figure 2: Left: The functional $I_1(\mathbf{w})$ is approximated in the vicinity of $\mathbf{w}^{(n)}$ by a surrogate functional $I_1^{(n)}(\mathbf{w})$, constructed in such a way that its minimum is easy to find (eq. (5)). This defines the next step in the iteration. Right: Soft thresholding function $\mathcal{S}_\tau(w)$.

is minimized (the crucial difference with $I_1(\mathbf{w})$ being the second term). This gives rise to the familiar system of damped normal equations

$$(\mathbf{A}^T\mathbf{A} + \tau\mathbf{I})\mathbf{m} = \mathbf{A}^T\mathbf{d}, \tag{10}$$

whose solution $\mathbf{m} = (\mathbf{A}^T\mathbf{A} + \tau\mathbf{I})^{-1}\mathbf{A}^T\mathbf{d}$ can be found using a linear solver of choice, since $\mathbf{A}^T\mathbf{A} + \tau\mathbf{I}$ is a regular matrix. To emphasize the similarities and differences with the ℓ_1 method, we adopt the classical Landweber iteration [Landweber(1951)] that can be (but in modern applications seldom is) used for solving the linear equations (10):

$$\mathbf{m}^{(n+1)} = \mathbf{A}^T\mathbf{d} + \left[\mathbf{I} - (\mathbf{A}^T\mathbf{A} + \tau\mathbf{I})\right]\mathbf{m}^{(n)}, \qquad \mathbf{m}^{(0)} = \mathbf{0}. \tag{11}$$

No thresholding is employed here. Rescaling of the operator and the data again improves the rate of convergence:

$$\mathbf{m}^{(n+1)} = \alpha^2\mathbf{A}^T\mathbf{d} + \left[\mathbf{I} - (\alpha^2\mathbf{A}^T\mathbf{A} + \tau\alpha^2\mathbf{I})\right]\mathbf{m}^{(n)}, \qquad \mathbf{m}^{(0)} = \mathbf{0}. \tag{12}$$

Of course it is also possible to solve the linear system (10) using a conjugate-gradient or similar algorithm in much less time.

A third option is to use an ℓ_2 penalization on the wavelet coefficients. This allows us to penalize the scaling coefficients differently than the wavelet coefficients (with the help of different penalization parameters τ_s and τ_w). We can use the following iteration, similar to formula (12), but now in the wavelet domain:

$$\mathbf{w}^{(n+1)} = \alpha^2\mathbf{W}\mathbf{A}^T\mathbf{d} + \left[\mathbf{I} - (\alpha^2\mathbf{W}\mathbf{A}^T\mathbf{A}\mathbf{W}^T + \alpha^2\tilde{\mathbf{I}})\right]\mathbf{w}^{(n)}, \qquad \mathbf{w}^{(0)} = \mathbf{0}, \tag{13}$$

where $\tilde{\mathbf{I}}$ acts as the $\tau_\mathrm{w} \times$ identity on wavelet coefficients and as the $\tau_\mathrm{s} \times$ identity on scaling coefficients. If we were to use an orthonormal wavelet basis ($\mathbf{W}^T\mathbf{W} = \mathbf{W}\mathbf{W}^T = \mathbf{I}$) for our expansions, and if we penalized every coefficient the same, $\tau_\mathrm{w} = \tau_\mathrm{s} = \tau$, then this method would be identical to the previous ℓ_2 method.

In the following we consider both one-step and two-step ℓ_1 wavelet penalization as well as conventional ℓ_2 penalization, both without and with wavelets, using the rescaled iterative schemes (7), (8), (12) and (13) for the purposes of comparison.

3 Implementation

To test the above ideas, we devised a dramatically simplified, two-dimensional, synthetic surface-wave inversion problem very loosely modeled after an actual PASSCAL deployment in Tanzania [Owens et al.(1995)]. Fig. 4 (left) shows the hypothetical experimental setup: the highly schematized input model consists of a sharp, bent, East African rift structure with low shear-wave velocity, $\delta \ln\beta(x,y) < 0$, superimposed upon a smooth, circular cratonic positive anomaly, $\delta \ln\beta(x,y) > 0$. Eleven earthquake events (circles) were taken from the NEIC catalogue to mimic realistic regional seismicity for the duration of a typical temporary deployment of the twenty-one stations (triangles). The locations of the seismic stations and events are listed in Table 1. For each of the 11×21 source-receiver paths, we assume that fundamental-mode Rayleigh-wave perturbations $\delta k(\nu)$ have been measured at eight selected frequencies between $\nu \approx 0.01$ Hz and $\nu \approx 0.1$ Hz. These wavenumber perturbations are related to the 2D, depth-independent velocity perturbations $\delta \ln\beta(x,y)$ via a 2D, frequency-dependent sensitivity kernel (see Appendix A for more details):

$$\delta k(\nu) = \iint K_{2D}(x,y,\nu)\, \delta \ln\beta(x,y)\, dx\, dy. \tag{14}$$

Plots of the lowest-frequency ($\nu \approx 0.01$ Hz) and highest-frequency ($\nu \approx 0.1$ Hz) kernel $K_{2D}(x,y,\nu)$ for a typical source-receiver pair are shown in the left two panels of Fig. 5. Because finite-frequency scattering and diffraction effects are accounted for in the kernels $K_{2D}(x,y,\nu)$, there is significant off-path sensitivity of the measurements $\delta k(\nu)$ within the first one or two Fresnel zones [Zhou et al.(2004)]. All kernels $K_{2D}(x,y,\nu)$ and distances are computed in the flat-earth earth approximation.

The study region, which is 35° (north-south) by 25° (east-west), is subdivided into $N_x \times N_y = 64 \times 64 = 4096$ equal-sized rectangles, and the discretized model vector $\mathbf{m}$ consists of the unknown constant values of $\delta \ln\beta(x,y)$ within each rectangle. To compute the matrix $\mathbf{A}$, which maps the discretized model $\mathbf{m}$ onto the data $\mathbf{d}$ (consisting of multiple $\delta k(\nu)$), each kernel $K_{2D}(x,y,\nu)$ is sampled $n_x \times n_y$ times on each of the $N_x \times N_y$ model-vector rectangles and a Riemann sum is used to compute the quantity

$$\iint_{\text{rectangle}(k,l)} K_{2D}(x,y,\nu)\, dx\, dy \approx \frac{\Delta x \Delta y}{n_x n_y} \sum_{m,n} K_{2D}\left(x_l - \Delta x/2 + (m-1/2)\delta x,\, y_k - \Delta y/2 + (n-1/2)\delta y,\, \nu\right),$$
$$\tag{15}$$

where $\delta x = \Delta x/n_x$ and $\delta y = \Delta y/n_y$. A schematic representation of the $N_x \times N_y$ grid on which the spatial-domain model $\mathbf{m}$ is specified and the $n_x \times n_y$ integration subgrid is shown in Fig. 3. We choose $n_x = n_y = 32$ since we have found that doubling this to $n_x = n_y = 64$ yields a change of less than one percent in the integrated value of $\mathbf{A}$. The dimensions of the resulting matrix $\mathbf{A}$ are 1848 (number of stations × number of events × number of wavenumbers) by 4096 (number of model-vector pixels). To give an idea of the overall degree of coverage, we have plotted the sum (over all station event pairs) of the absolute value of all of the lowest-frequency and all the highest-frequency discretized kernels in the right two panels of Fig. 5. It is clear that much of the study area, particularly in the northwest and southeast, is completely uncovered (as is typical of real-world, regional seismic experiments).

Using the matrix $\mathbf{A}$ and the input model $\mathbf{m}^{\text{input}}$ with a sharp, low-velocity East African rift superimposed on a broad, high-velocity cratonic structure, we compute synthetic data $\mathbf{d} = \mathbf{A}\mathbf{m}^{\text{input}} + \mathbf{e}$, where we have added Gaussian noise $\mathbf{e}$ with zero mean and a standard deviation equal to two percent of the largest synthetic wavenumber perturbation, i.e. $\sigma = 0.02 \max(|\mathbf{A}\mathbf{m}^{\text{input}}|)$. By adopting a constant standard deviation σ, errors at the highest frequency ν and unperturbed wavenumber $k(\nu)$ are more than an order of magnitude smaller than those for the lowest frequency and wavenumber data, where the signal-to-noise ratio may be close to unity. Since finite-frequency inversions include the effect of scattered wave energy, a high precision of the measurement $\delta k(\nu)$ at high frequency ν is realistic. The purpose of the proposed algorithm is now to reconstruct $\mathbf{m}$ from the knowledge of the noisy data $\mathbf{d}$, the matrix $\mathbf{A}$ and the linear equations $\mathbf{A}\mathbf{m} = \mathbf{d}$.
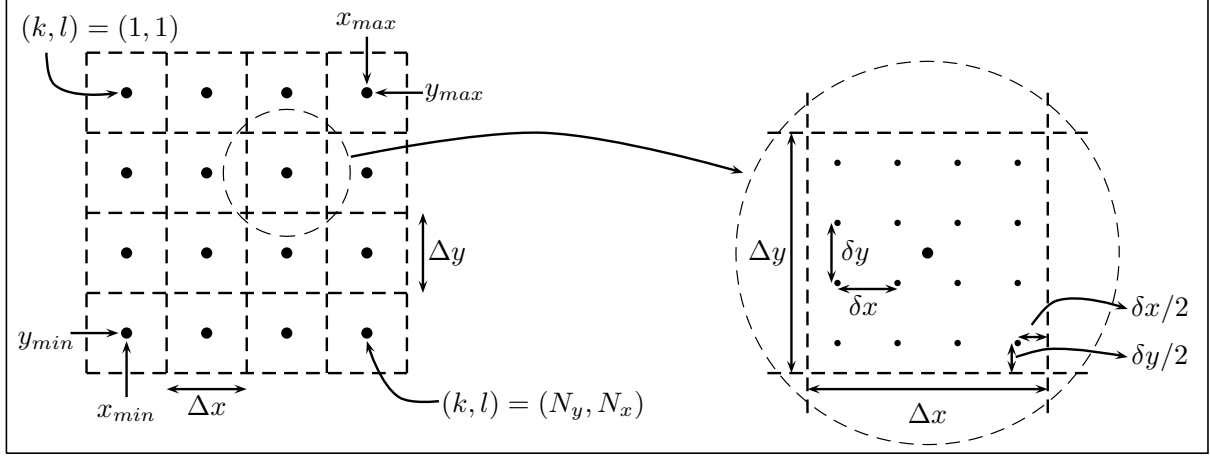
Figure 3: Schematic representation of the 2D Cartesian grids used. Left: $N_x \times N_y$ grid used to specify the model $\mathbf{m}$. Right: Blowup of the finer-scale $n_x \times n_y$ grid used to compute the kernel matrix $\mathbf{A}$ via the approximate integration (15). Since the study region is rectangular in shape (see Fig. 4) and since $N_x = N_y$ and $n_x = n_y$, the actual $\Delta x \times \Delta y$ model pixels and $\delta x \times \delta y$ integration subpixels are also rectangular, rather than square as shown.

Table 1: List of positions of seismic stations and earthquake events used in the synthetic inversion.

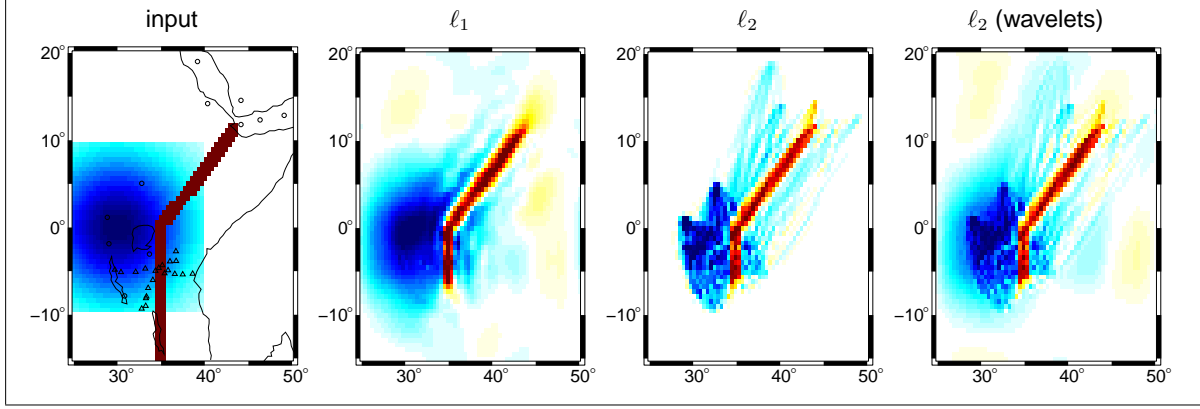| Stations		Events	
longitude	latitude	longitude	latitude
33.3203°	−7.9073°	29.02°	−1.86°
35.1382°	−4.3238°	49.10°	12.85°
32.7712°	−9.2958°	44.15°	11.80°
33.2588°	−8.1060°	30.82°	−7.84°
29.6927°	−4.8392°	46.34°	12.33°
38.6170°	−5.3018°	39.17°	19.02°
30.3988°	−5.1168°	32.78°	5.06°
36.5695°	−5.3223°	44.15°	14.57°
37.4763°	−5.3775°	28.84°	1.16°
36.7192°	−3.8422°	40.33°	14.20°
35.7965°	−4.9040°	33.67°	−3.05°
36.6983°	−2.7252°		
34.3462°	−4.9610°		
34.0560°	−6.0192°		
35.4007°	−5.2508°		
33.2415°	−8.9835°		
33.1842°	−4.7145°		
33.5180°	−6.9372°		
34.7315°	−4.6403°		
36.0163°	−3.8892°		
32.0832°	−5.0878°		

Figure 4: From left to right: Toy 2D velocity model for the East African rift and adjacent continental craton, showing the seismic stations (triangles) and earthquake events (circles); reconstructed model using the two-step ℓ_1-penalization method; reconstruction using the spatial-domain ℓ_2 method; reconstruction using the wavelet-domain ℓ_2 method. The two-step ℓ_1 model is that obtained after $1000+1000$ iterations, whereas both ℓ_2 models are after 2000 Landweber iterations. Red denotes low anomalous velocity, $\delta\ln\beta(x,y) < 0$, and blue denotes high velocity, $\delta\ln\beta(x,y) > 0$. The absolute magnitude $|\delta\ln\beta(x,y)|$ is irrelevant, since the inverse problem $\mathbf{Am} = \mathbf{d}$ is linear and the synthetic data are constructed from the input model $\mathbf{m}^{\text{input}}$ (leftmost map) via $\mathbf{d} = \mathbf{Am}^{\text{input}} + \mathbf{e}$.
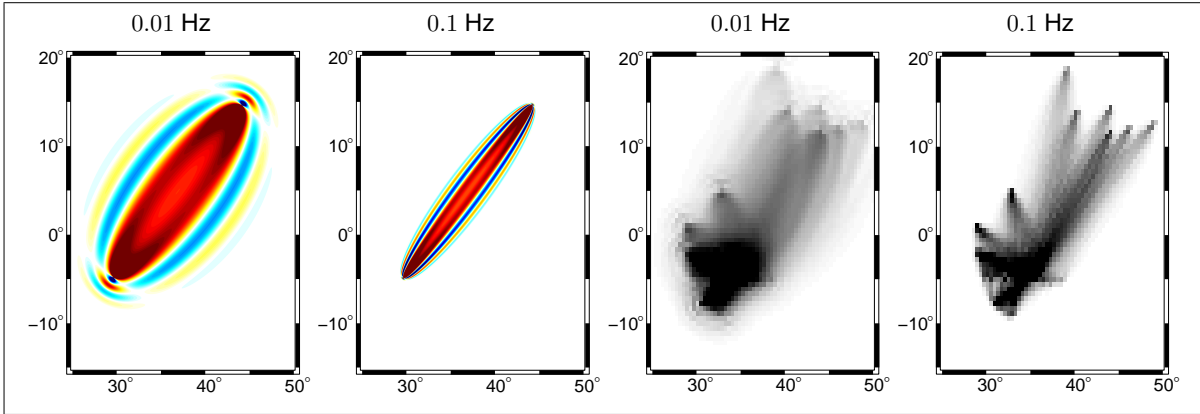

Figure 5: Left: Map view of a typical two-dimensional sensitivity kernel $K_{2D}(x,y,\nu)$ at the lowest frequency considered, $\nu \approx 0.01$ Hz. Second from left: Highest-frequency ($\nu \approx 0.1$ Hz) kernel for the same source-receiver path. Both kernels exhibit structure on a much finer scale than the resolution of the model, necessitating the $n_x \times n_y$ numerical integration to compute the matrix $\mathbf{A}$ in eq. (15). Red denotes negative values, $K_{2D}(x,y,\nu) < 0$, and blue denotes positive values, $K_{2D}(x,y,\nu) > 0$. The cross-path tapering of the kernels as a result of the finite time-domain taper, eqs (A6)–(A7), is clearly visible. Second from right: The sum (over all source-receiver pairs) of the absolute value of the lowest frequency ($\nu \approx 0.01$ Hz) integrated kernels (as computed in eq. 15). Far right: The sum (over all source-receiver pairs) of the absolute value of the highest frequency ($\nu \approx 0.1$ Hz) integrated kernels (as computed in eq. 15). The coverage is adequate in the vicinity of the East African rift (by design of the original seismic deployment) but poor elsewhere.
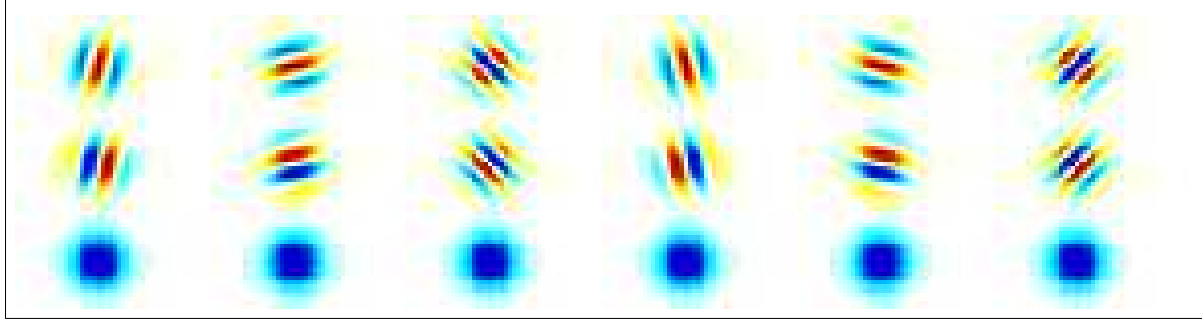
Figure 6: Spatial-domain structure of the 2D dual-tree complex wavelets used in the reconstruction (figure taken/adapted from [Selesnick et al.(2006)]). First row: real part, second row: imaginary part, third row: norm squared (i.e. sum of the squares of the top two plots). The directional character of each of the six wavelet functions is clear. Four different wavelet scales, i.e. four different replicas of this picture, each a factor of two smaller than the one above it, are used in both the ℓ_1 and ℓ_2 wavelet-basis inversions.

For our purposes we will make use of the overcomplete 2D wavelet basis described by [Kingsbury(2002)] and [Selesnick et al.(2006)] because of its ability to distinguish different directions (see Fig. 6). We use four wavelet scales, for a total of $4 \times 64^2 = 16\,384$ wavelet and scaling coefficients $\mathbf{w}$ (four times the number of model coefficients $\mathbf{m}$). The starting point for the iterations in both the ℓ_1 and ℓ_2 inversions is $\mathbf{w} = \mathbf{0}$ and $\mathbf{m} = \mathbf{0}$. As explained in the previous section we renormalize the ℓ_1 thresholded iteration by choosing $\alpha = \lambda_{\max}^{-1/2}$ (which in our case equals 4884.5) where $\lambda_{\max}$ is the largest eigenvalue of $\mathbf{A}^T\mathbf{A}$. We let the iteration run for 1000 steps, adjust the data (two-step procedure) and let the second-step algorithm run for another 1000 steps. The threshold τ is chosen by hand in such a way as to arrive at a final value for the variance-adjusted misfit, $\chi^2 = \|\mathbf{d} - \mathbf{Am}\|_2^2/\sigma^2$, that is approximately equal to 1848 (the number of data). The noisy data $\mathbf{d}$ are thus fit to within their standard errors σ and no better; pushing the fit beyond this would amount to fitting the noise $\mathbf{e}$, which would lead to undesirable artifacts in the resulting model $\mathbf{m}$.

It should also be noted that the thresholding is done on pairs of wavelet coefficients: The wavelets come in pairs (at the same scale, position and orientation) that we interpret as real and imaginary part of a complex wavelet, i.e. thresholding corresponds to $(w_k^{\mathrm{re}}, w_k^{\mathrm{im}}) \to z = w_k^{\mathrm{re}} + iw_k^{\mathrm{im}} \to \tilde{z} = z\mathrm{S}_\tau(|z|)/|z| \to (\mathrm{Re}(\tilde{z}), \mathrm{Im}(\tilde{z}))$. This particular method of thresholding is borrowed from image denoising where it is found to make a big difference in avoiding artifacts [Guleryuz(2006), van Spaendonck et al.(2003), Selesnick et al.(2005)]. Furthermore, the threshold for the diagonally oriented wavelets is multiplied by 1.2395 because $\|\nabla\psi_{\pm 45^\circ}\|_1 = 1.2395\|\nabla\psi_{\mathrm{other}}\|_1$. We choose the threshold τ_{s} for the scaling coefficients to be 1/10th of the threshold τ_{w} for the wavelet coefficients; since the scaling coefficients correspond to a few large-scale averages (64 in our case versus more than 16 000 finer-scale wavelet coefficients) it is not so important that these be sparse. Likewise, in the wavelet-basis ℓ_2 inversions, we set the penalization parameter for the scaling coefficients to 1/10th the value of the penalization parameter for the wavelet coefficients.

The two-step ℓ_1 algorithm takes about ten minutes for $1000 + 1000$ iterations on a 1.5GHz PC. The result of the ℓ_1 inversion is compared with the outcome of both of the ℓ_2 methods, with and without using wavelets, with the thresholding or penalization parameter τ chosen in every case to achieve the same data fit: $\chi^2 \approx 1848$ (see Fig. 7). The number of Landweber iterations is 2000, so that the total number of two-step ℓ_1 and single-step ℓ_2 iterations is the same. The spatial-domain ℓ_2-regularization method yields a relative modeling error $\|\mathbf{m} - \mathbf{m}^{\mathrm{input}}\|_2/\|\mathbf{m}^{\mathrm{input}}\|_2$ of about 74%, whereas the two-step ℓ_1 method yields a relative modeling error of only 47% (see Fig. 8), and is clearly less noisy (compare the

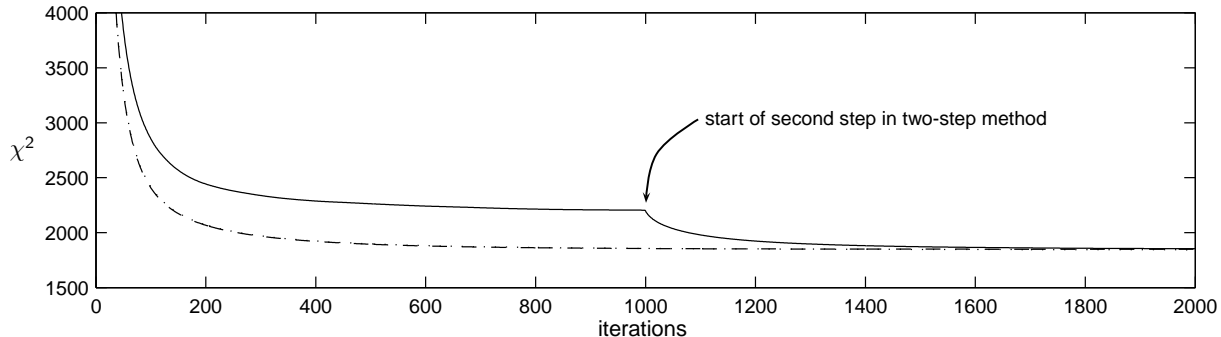Figure 7: Graph of the variance-adjusted data fit $\chi^2 = \|\mathbf{d} - \mathbf{Am}\|_2^2 / \sigma^2$ versus the number of iterations: two-step ℓ_1-regularization method (solid line), spatial-domain ℓ_2 method (dashed line) and wavelet-basis ℓ_2 method (dotted line). The thresholding and penalization parameter τ has in each case been tailored so that the final value of χ^2, after $1000 + 1000$ or 2000 iterations, is equal to the number of data, namely 1848. Note the improvement in the rate of convergence toward the model with $\chi^2 = 1848$ after the implementation of the second step in the two-step ℓ_1 iteration.

middle two maps in Fig. 4). The wavelet-basis ℓ_2-regularized inversion (rightmost map in Fig. 4) is only slightly less noisy, with a relative modelling error of about 55% (Fig. 8). One feature that can never be recovered in any of the reconstructions is the southern part of the rift, which does not lie between any station-event pair.

In Fig. 9 we compare the wavelet coefficients $\mathbf{w}$ of the input model, the two-step ℓ_1 reconstruction and the wavelet-basis ℓ_2 reconstruction. In accordance with our basic assumption, the ℓ_1-regularized model is sparse in the wavelet basis. Most of the small-scale coefficients $\mathbf{w}$ are zero —in agreement with the original model on the left— indicating the effectiveness of the iterative thresholding algorithm. The wavelet coefficients of the wavelet ℓ_2 reconstruction are clearly not sparse. Also this solution seems to suffer from large-scale artifacts (see Fig. 4, rightmost map).

In the leftmost plot in Fig. 10 we show χ^2 versus $\|\mathbf{w}\|_1$ tradeoff curves for the ℓ_1 reconstruction method, both with and without using the two-step procedure. After $1000 + 1000$ iterations, the ℓ_1 wavelet norm $\|\mathbf{w}\|_1$ of the two-step reconstructed model is lower — for the same value of χ^2 — than the corresponding norm of the model produced by 2000 iterations of the first step, with no subsequent redefinition of the data $\mathbf{d}$ and reiteration. This is an indication that 2000 total iterations is inadequate to achieve full convergence, since the fully converged model, which minimizes the functional $I_1(\mathbf{w})$ given in eq. (1), must be the minimum-norm model for a fixed value of χ^2 by definition. A much larger number of iterations seems to be required to guarantee convergence. To construct the second set of tradeoff curves in Fig. 10, we employed $150\,000 + 150\,000$ iterations in the two-step case and $300\,000$ in the single-step case; such a large number would be prohibitive in any larger-scale, more realistic, 3D application. We have chosen to limit the iteration counts to $1000 + 1000$ or 2000 in all of our model-space comparisons, since any changes in the spatial-domain features of the models $\mathbf{m}$ are barely discernible to the eye with further iteration. The rightmost plot in Fig. 10 shows the principal advantage of using the two-step iteration procedure: for the same total number of iterations, either $1000 + 1000 = 2000$ or $150\,000 + 150\,000 = 300\,000$, the number of nonzero wavelet coefficients of the two-step models is always lower than the corresponding number for the single-step models. The two-step ℓ_1 procedure therefore leads more quickly to a sparser wavelet-basis solution, as expected.

We also compared the single-step and two-step ℓ_1 inversion methods with the corresponding ℓ_2 reconstruction methods, both with and without wavelets, for a number of other input synthetic models. These include three checkerboard patterns of decreasing scale and a model similar to the geologically
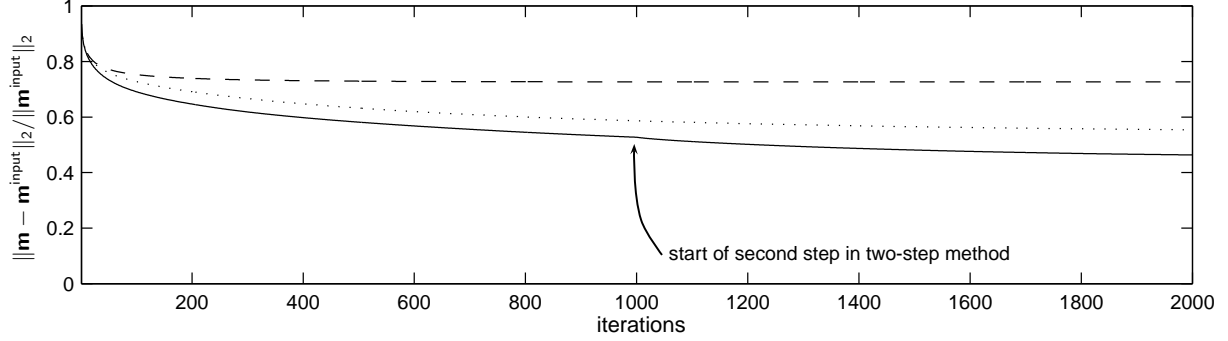
10

Figure 8: Graph showing the relative modeling error $\|\mathbf{m} - \mathbf{m}^{\mathrm{input}}\|_2/\|\mathbf{m}^{\mathrm{input}}\|_2$ versus the number of iterations: two-step ℓ_1 method (solid line), spatial-domain ℓ_2 method (dashed line) and wavelet-basis ℓ_2 method (dotted line). The ℓ_1-regularization method clearly yields the most faithful reconstruction of the input model $\mathbf{m}^{\mathrm{input}}$. Note the (slight) improvement in the rate of decrease of the modelling error following the start of the second step in the two-step ℓ_1 iteration.
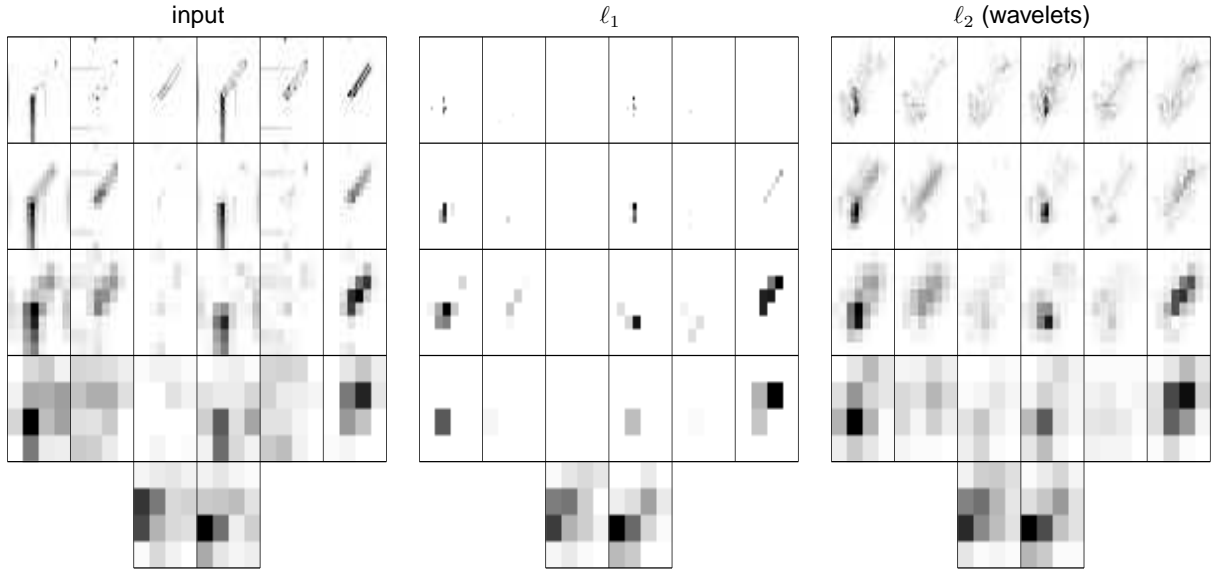

Figure 9: Graphical display of (the modulus of) the wavelet and scaling coefficients. Left: coefficients of the synthetic input model $\mathbf{m}^{\mathrm{input}}$. Middle: coefficients of the two-step ℓ_1 model after $1000 + 1000$ iterations. Right: coefficients of the wavelet-basis ℓ_2 model after 2000 iterations. The four wavelet scales are plotted, smallest to largest, top to bottom. Each row shows the six different wavelet directions, plotted next to each other in the same left-to-right order as the wavelets plotted in Fig. 6. Scaling coefficients are plotted on the bottom row. White denotes a zero coefficient, $w_i = 0$. Each rectangle corresponds to the spatial domain 25°E – 50°E by 15°S – 20°N.
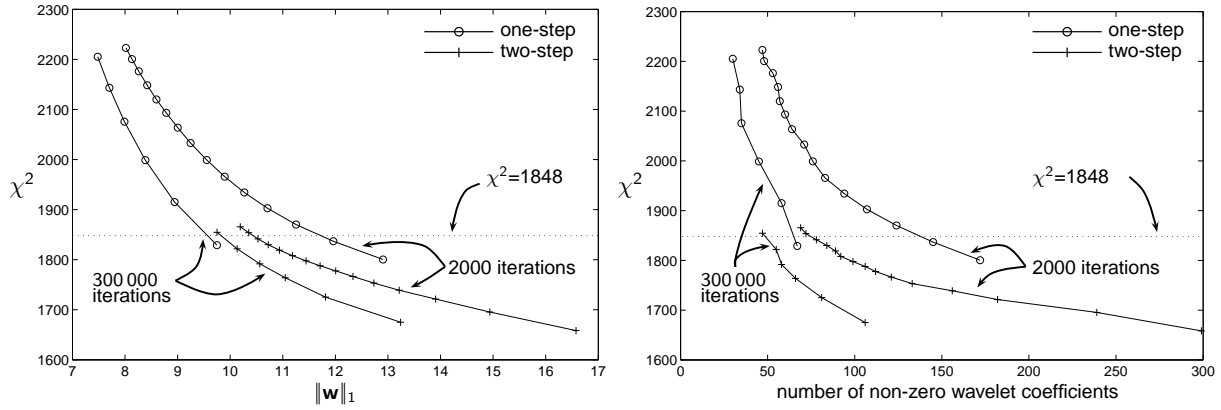
11

Figure 10: Left: Data misfit χ^2 versus ℓ_1 wavelet norm $||\mathbf{w}||_1$ tradeoff curve. Right: Alternative tradeoff curve showing χ^2 versus the number of nonzero wavelet coefficients of the model $\mathbf{m}$. Different values of the thresholding parameter τ were used to determine each point on the various curves. Circles: one-step method, after either 2000 or 300 000 iterations; crosses: two-step method after an equivalent number (either 1000 + 1000 or 150 000 + 150 000) of total iterations. The relative positions of the one-step and two-step curves suggests that 300 000 total iterations is sufficient to achieve full convergence. The horizontal dotted lines show the statistically meaningful value of the noisy data misfit, $\chi^2 = 1848$ (the number of data).

inspired one in Fig. 4, but with a more curvaceous low-velocity rift (see Fig. 11). Both the single-step ℓ_1 reconstructions and the ℓ_2 reconstructions are computed using 2000 iterations, whereas the two-step ℓ_1 models are computed using 1000 + 1000 iterations. In all cases, the two-step ℓ_1 models are the most parsimonious and therefore to most geoscientists the most acceptable. One could consider using smoothness damping to improve the quality of the ℓ_2 images; however, this would be done at the cost of resolving the sharpness of the rift structure. A nitpicker could perhaps also argue that the "rift" structure in the model produced by the ℓ_1 procedure extends further northwards, albeit diminished in amplitude, whereas conventional ℓ_2 regularization without wavelets exhibits a sharper cutoff, more like the input model. It achieves this sharp cutoff, however, at the expense of many artifacts elsewhere, especially along dominant ray directions. Perhaps the most noteworthy feature of the ℓ_1 regularization method is its suppression of the artifacts resembling high-frequency kernel images that are streaked along surface-wave raypaths in all the ℓ_2 models, to the north of the rift and within the craton. This is one of the most serious artifacts that plague conventional seismic tomography: ℓ_2 regularization frequently if not always seems to enhance the well-sampled regions of the model. The ℓ_1 wavelet-basis reconstructions show no signs of this familiar deficiency.

The computational bottleneck in the present 2D synthetic study is not the wavelet transform — which is fast, certainly on a model $\mathbf{m}$ of modest dimension 64×64 — or even the number of iterations, but it is simply the size of the matrix $\mathbf{A}$. A significant amount of time is needed to accurately pre-compute $\mathbf{A}$, and considerable memory is needed to store the computed elements in memory; this is necessary because the product $\mathbf{A}^T\mathbf{A}$ is used in every step of the iteration. Doubling of the resolution in every direction results in a fourfold increase in size of the model $\mathbf{m}$, and a sixteen-fold increase in the number of elements in the square matrix $\mathbf{A}^T\mathbf{A}$. All calculations were performed using Matlab; software for the 2D dual-tree wavelets was downloaded from [Selesnick et al.(2006)].
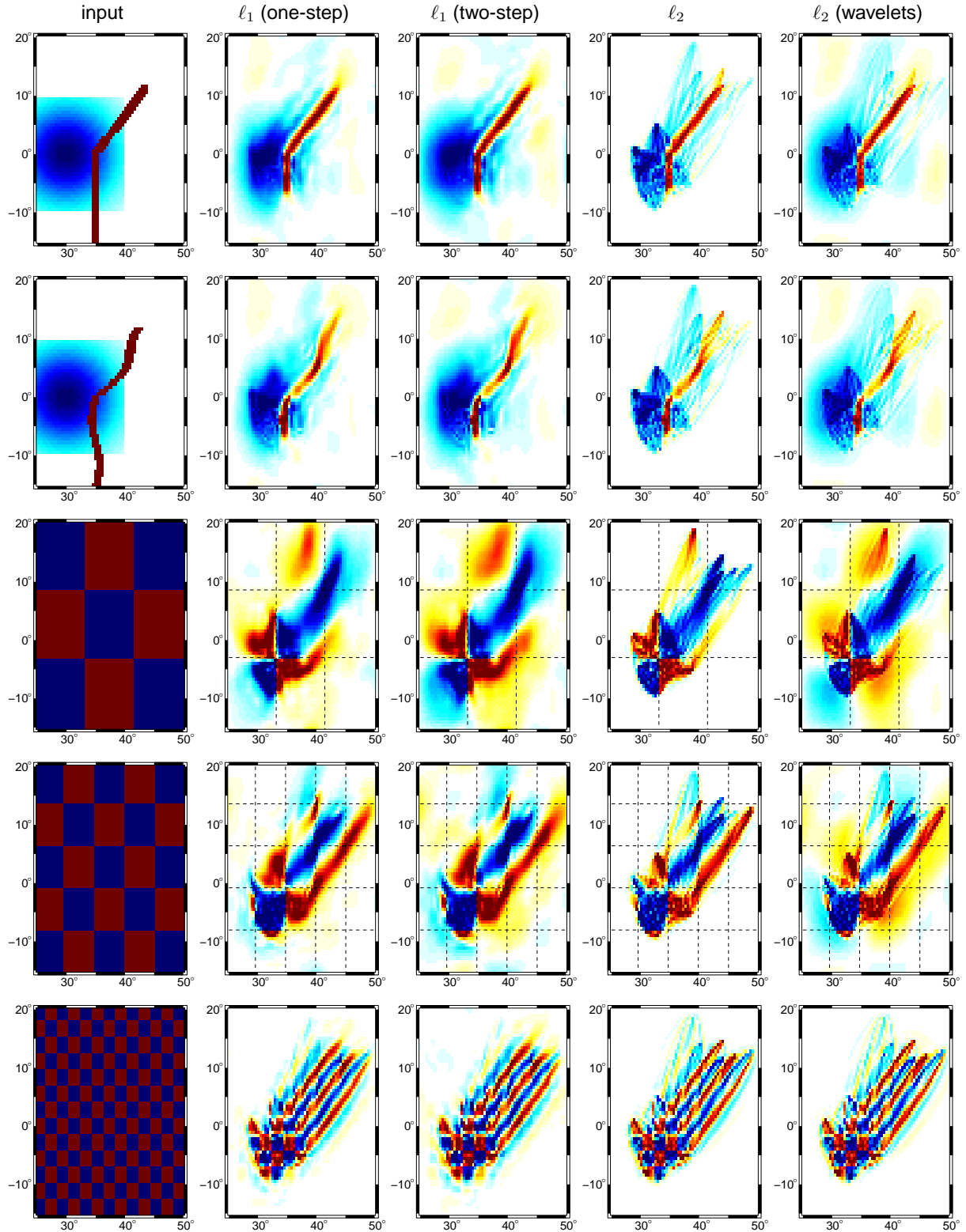
Figure 11: Results of applying different reconstruction techniques to a number of different 2D toy models. From left to right: Original input model; ℓ_1 reconstruction (2000 iterations, single-step procedure); ℓ_1 reconstruction (1000 + 1000 iterations, two–step procedure); ℓ_2 reconstruction (2000 iterations, without wavelets); ℓ_2 reconstruction (2000 iterations, with wavelets).

4 Conclusions

We tested several new methods of regularization through wavelet decomposition of a toy 2D tomographic problem characterized by both smooth and sharp velocity anomalies. A variety of synthetic inversion experiments show that minimization of the ℓ_1-norm of a wavelet decomposition of the model leads to tomographic images that are parsimonious in the sense that they use only a few wavelets and still represent both smooth and sharp features well without introducing significant blurring or artifacts. The ℓ_1-norm performs significantly better than an ℓ_2 regularization on either the model or its wavelet decomposition. In particular, raypath-associated artifacts are almost completely suppressed.

The choice of dual-tree complex wavelets in 2D, representing six space directions, is sufficient to avoid directional bias, and efficient in modeling both smooth features such as the cratonic structure as well as sharp features such as the rift structure in our simplified synthetic model. Numerical comparisons between the inversion results and the input model used to generate the data confirm the superiority of the ℓ_1-norm regularization. Though in real-world inversions such ground-truth information is not available, one can argue that the ℓ_1 inversion method serves the principle of parsimony well and is to be preferred over more common methods. If the tomographic object (such as the real earth) is too complex to be well represented by a parsimonious expansion in wavelets, neither method is able to resolve such complexity adequately with a limited data set, as shown in the bottom rows of Fig. 11, where even the ℓ_1 inversions begin to show the effects of raypath distribution. In this case, we expect that the principle of parsimony can be usefully applied once a richer family of building blocks is considered.

The only drawback of the method, so far, is the slow convergence of the ℓ_1 surrogate-functional iteration procedure. Our preference for the thresholded algorithm used here arises from the fact that its convergence is guaranteed even though the ℓ_1 problem is nonlinear. We have introduced a two-step procedure that leads to a significant speedup; however, Fig. 10 indicates that even 1000+1000 iterations do not suffice for complete convergence (it nevertheless produces an excellent approximation). A potentially promising approach towards further convergence improvement is to combine an efficient linear method (such as e.g. conjugate-gradient) with an *adaptive* thresholding scheme. This would then avoid the need to precompute the largest eigenvalue of $\mathbf{A}^T\mathbf{A}$ and facilitate the application of the ℓ_1 method to a larger, 3D, study of body-wave tomography.

5 Acknowledgments

References

[Candes et al.(2006)] Candes, E., Romberg, J. & Tao, T., 2006. Stable signal recovery from incomplete and inaccurate measurements, to appear in *Comm. Pure Appl. Math.*

[Constable et al.(1987)] Constable, S.C., Parker, R.L. & Constable, C.G., 1987. Occam's inversion: a practical algorithm for generating smooth models from electromagnetic sounding data, *Geophys.*, **52**, 289–300.

[Daubechies(1992)] Daubechies, I., 1992. *Ten Lectures on Wavelets*, SIAM Press, Philadelphia.

[Daubechies et al.(2004)] Daubechies, I., Defrise, M. & De Mol, C., 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure App. Math.*, **57**, 1413–1541, arXiv/FA/0307152.

[Deal & Nolet(1999)] Deal, M.M. & Nolet, G., 1999. Slab temperature and thickness from seismic tomography 2. Izu-Bonin, Japan and Kuril subduction zones, *J. Geophys. Res.*, **104**, 28803–28812.

[Deal et al.(1999)] Deal, M.M., Nolet, G. & van der Hilst, R.D., 1999. Slab temperature and thickness from seismic tomography, 1. Method and application to Tonga, *J. Geophys. Res.*, **104**, 28789–28802.

[de Hoop & van der Hilst(2005)] de Hoop, M. V. & van der Hilst, R. D., 2005. On sensitivity kernels for wave equation tomography, *Geophys. J. Int.*, **160**, 621–633.

[Donoho(2004)] Donoho, D. L., 2006. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution, *Comm. Pure Appl. Math.*, **59**, 797–829.

[Dziewonski et al.(1975)] Dziewonski, A.M., Hager, B.H. & O'Connell, R.J., 1975. Large scale heterogeneities in the lower mantle, *J. Geophys. Res.*, **82**, 239–255.

[Dziewonski & Woodhouse(1987)] Dziewonski, A.M. & Woodhouse, J.H., 1987. Global images of the Earth interior, *Science*, **236**, 37–48.

[Guleryuz(2006)] Guleryuz, O.G, 2003. Weighted overcomplete denoising, In *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, **2**, 1992–1996.

[Kingsbury(1999)] Kingsbury, N., 1999. Image processing with complex wavelets, *Phil. Trans. Roy. Soc. Lond.*, **A357**, 2543–2560.

[Kingsbury(2002)] Kingsbury, N. G., 2002. Complex wavelets for shift invariant analysis and filtering of signals, *Applied and Computational Harmonic Analysis*, **10**, 234–253.

[Landweber(1951)] Landweber, L., 1951. An iterative formula for Fredholm integral equations of the first kind, *Am. J. Math.*, **73**, 615–624

[Masters et al.(1996)] Masters, G., Johnson, S., Laske, G. & Bolton, H., 1996. A shear veolcity model of the mantle, *Phil. Trans. Roy. Soc. Lond.*, **A354**, 1385–1410.

[Nolet(1987)] Nolet, G., 1987. Seismic wave propagation and seismic tomography, In G. Nolet, editor, *Seismic Tomography*, pages 1–23, Dordrecht, Reidel.

[Owens et al.(1995)] Owens, T.J., Nyblade, A.A. & Langston, C.A., 1995. The Tanzania broadband experiment, *IRIS Newsletter*, **14**, 1.

[Portniaguine & Zhdanov(1999)] Portniaguine, O. & Zhdanov, M.S., 1999. Focusing geophysical inversion images, *Geophys.*, **64**, 874–887.

[Selesnick et al.(2005)] Selesnick, I. W., Baraniuk, R. G. & Kingsbury, N., 2005. The dual-tree complex wavelet transform — A coherent framework for multiscale signal and image processing, *IEEE Signal Processing Magazine*, **22**, 123-151.

[Selesnick et al.(2006)] Selesnick, I., Cai, S. & Li, K., 2006. MATLAB implementation of wavelet transforms, `http://taco.poly.edu/WaveletSoftware/`.

[Spakman & Nolet(1988)] Spakman, W. & Nolet, G., 1988. Imaging algorithms, accuracy and resolution in delay-time tomography, in N.J. Vlaar et al., editor, *Mathematical Geophysics*, pages 155–187, Hingham, Mass., Reidel.

[Trampert & Snieder(1996)] Trampert, J. & Snieder, R., 1996. Model estimations biased by truncated expansions: possible artifacts in seismic tomography, *Science*, **271**, 1257–1260.
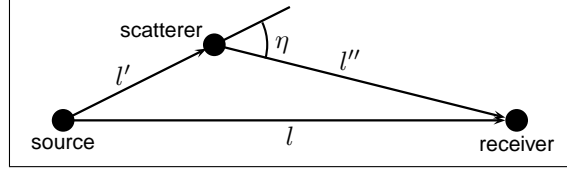
Figure 12: Schematic map view of the single-scattering geometry in our simplified 2D, flat-earth, surface-wave inversion problem. The quantity l is the horizontal epicentral distance between the source and receiver; l' and l'' are the lengths of the first and second legs of the detour path, respectively, and the angle η measures the deflection of the wave at the scatterer.

[VanDecar & Snieder(1994)] VanDecar, J.C. & Snieder, R., 1994. Obtaining smooth solutions to large, linear, inverse problems, *Geophysics*, **59**, 818–829.

[van Spaendonck et al.(2003)] van Spaendonck, R., Blu, T., Baraniuk, R. & Vetterli, M., 2003. Orthogonal Hilbert transform filter banks and wavelets, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, **6**, 505-508.

[Zhou et al.(2004)] Zhou, Y., Dahlen, F.A. & Nolet, G., 2004. Three-dimensional sensitivity kernels for surface wave observables, *Geophys. J. Int.*, **158**, 142–168.

A   Two-dimensional sensitivity kernels

The toy linear inverse problem $\mathbf{Am} = \mathbf{d}$ used in this paper is designed to incorporate all the important characteristics of a real-world regional tomographic inversion, while at the same time being small enough to allow for repeated experimenting with reasonable CPU times on a single workstation. For this reason, we limit attention to surface-wave dispersion data, specifically perturbations $\delta k(\nu)$ in the wavenumber $k(\nu)$, presumed to be measured in rad/m, of the fundamental ($n = 0$) Rayleigh mode at temporal frequency ν, measured in Hz. Finite-frequency theory based upon the Born approximation [Zhou et al.(2004)] gives a linear relationship between such wavenumber perturbations and the 3D perturbations in the fractional shear-wave velocity $\delta\ln\beta(\mathbf{x})$ within the earth:

$$\delta k(\nu) = \iiint K_{\mathrm{3D}}(\mathbf{x}, \nu)\, \delta\ln\beta(\mathbf{x})\, \mathrm{d}^3\mathbf{x}. \tag{A1}$$

Making use of a number of flat-earth approximations that do not fundamentally affect the nature of the inverse problem, we can write the 3D Fréchet sensitivity kernel $K_{\mathrm{3D}}(\mathbf{x}, \nu)$, for the simplest case of an explosive source with an isotropic radiation pattern and a measurement made on the vertical component at the receiver, in the form

$$K_{\mathrm{3D}}(\mathbf{x}, \nu) = [e_0(z, \nu) + e_1(z, \nu)\cos\eta + e_2(z, \nu)\cos 2\eta]\left(\frac{1}{8\pi k(\nu)l\, l'l''}\right)^{\frac{1}{2}}\sin[k(\nu)(l' + l'' - l + \pi/4)], \tag{A2}$$

where z is the depth, l is the epicentral distance measured in m on the surface of the earth, and l' and l'' are the horizontal distances of the scatterer $\mathbf{x} = (x, y, z)$ from the source and receiver, respectively. The quantity η is the scattering angle, measured at the surface projection (x, y) of $\mathbf{x}$, as shown in Figure 12. Expressions for the depth-dependent functions $e_0(z, \nu)$, $e_1(z, \nu)$ and $e_2(z, \nu)$ can be found in the appendix of [Zhou et al.(2004)].

To simplify matters even further, we assume that the velocity perturbation $\delta\ln\beta(\mathbf{x})$ is independent of depth z and dependent only upon the horizontal Cartesian coordinates x and y. Upon integrating the factors $e_0(z,\nu)$, $e_1(z,\nu)$ and $e_2(z,\nu)$ over depth,

$$E_0(\nu) = \int_0^\infty e_0(z,\nu)\,\mathrm{d}z, \qquad E_1(\nu) = \int_0^\infty e_1(z,\nu)\,\mathrm{d}z, \qquad E_2(\nu) = \int_0^\infty e_2(z,\nu)\,\mathrm{d}z, \qquad \text{(A3)}$$

we may then relate $\delta k(\nu)$ to $\delta\ln\beta(x,y)$ via a 2D sensitivity kernel:

$$\delta k(\nu) = \iint K_{2\mathrm{D}}(x,y,\nu)\,\delta\ln\beta(x,y)\,\mathrm{d}x\,\mathrm{d}y, \qquad \text{(A4)}$$

where

$$K_{2\mathrm{D}}(x,y,\nu) = [E_0(\nu) + E_1(\nu)\cos\eta + E_2(\nu)\cos 2\eta]\left(\frac{1}{8\pi k(\nu)l\,l'l''}\right)^{\frac{1}{2}}\sin[k(\nu)(l'+l''-l+\pi/4)]. \quad \text{(A5)}$$

The rapidly oscillating sinusoidal function $\sin[k(\nu)(l'+l''-l+\pi/4)]$ in eq. (A5) is constant on ellipses, $l'+l'' = $ constant, having the surface projections of the source and receiver as foci. The $\cos\eta$ and $\cos 2\eta$ dependence and the term involving the integrable singularity $1/\sqrt{l'l''}$ act to slowly modulate this dominant elliptical dependence.

Eqs (A4) and (A5) are valid, subject to the already noted approximations, for a monochromatic wavenumber perturbation $\delta k(\nu)$, whereas actual surface-wave dispersion measurements must of necessity be made on a portion of a seismogram of finite length, typically multiplied by a time-domain taper $h(t)$. [Zhou et al.(2004)] show that the effect of such a finite-length taper can be accounted for by modifying the taper as follows:

$$K_{2\mathrm{D}}(x,y,\nu) \to K_{2\mathrm{D}}(x,y,\nu)\,h((l'+l'')/C(\nu)), \qquad \text{(A6)}$$

where $C(\nu)$ is the the group velocity at frequency ν measured in m/s. This modification has the effect of limiting the cross-path width of the Fréchet kernel $K(x,y,\nu)$, since $h(t) = 0$ for large detour times. We assume the data $\delta k(\nu)$ have been measured using a Hann or cosine taper, of duration five wave periods centered on the group arrival time:

$$h(t) = \begin{cases} 0 & \text{for} \quad t \leq t_{\mathrm{arrival}} - 2.5/\nu \\ \frac{1}{2}[1 - \cos 2\pi\nu(t - t_{\mathrm{arrival}} - 2.5/\nu)] & \text{for} \quad t_{\mathrm{arrival}} - 2.5/\nu \leq t \leq t_{\mathrm{arrival}} + 2.5/\nu \\ 0 & \text{for} \quad t \geq t_{\mathrm{arrival}} + 2.5/\nu \end{cases} \qquad \text{(A7)}$$

where $t_{\mathrm{arrival}} = l/C(\nu)$. Since $l'+l'' \geq l$ only the $t \geq t_{\mathrm{arrival}}$ portion of the taper (A7) contributes to the finite-record-length sensitivity kernel (A6).

The group velocity $C(\nu)$, unperturbed wavenumber $k(\nu)$ and auxiliary variables $E_0(\nu)$, $E_1(\nu)$ and $E_2(\nu)$ for fundamental-mode Rayleigh waves are listed in Table 2 at the eight selected frequencies ν; the corresponding wave periods vary roughly between 100 and 10 s. Since $E_0(\nu)$, $E_1(\nu)$ and $E_2(\nu)$ are all negative, a positive velocity perturbation, $\delta\ln\beta(x,y) > 0$, gives rise to a negative wavenumber perturbation, $\delta k(\nu) < 0$, i.e. an apparently longer wavelength wave, as expected. See Fig. 5 for two examples of sensitivity kernels $K_{2\mathrm{D}}(x,y,\nu)$ computed in this way. It is noteworthy that a 2D surface-wave inversion based upon eqs (A4)–(A7) differs from the common approach of inverting for a 2D phase velocity map at a single specified frequency ν: such maps are strictly incompatible with the notion of finite frequency, where no local phase velocity can be defined except when very crude approximations are made; for a discussion of this issue see [Zhou et al.(2004)].

B Notes on wavelets

The basic building block of the 1D discrete wavelet transform (DWT) is a filter bank. It consists of a high-pass filter g (i.e. a generalized difference) and a low-pass filter h (i.e. a generalized average) that

Table 2: Parameters $\nu, C(\nu), k(\nu), E_0(\nu), E_1(\nu)$ and $E_2(\nu)$ needed to compute the simplified 2D sensitivity kernels $K_{2D}(x, y, \nu)$. Fundamental-mode Rayleigh-wave measurements $\delta k(\nu)$ are presumed to have been made at eight frequencies ranging between $\nu \approx 0.01$ Hz (100 s period) and $\nu \approx 0.1$ Hz (10 s period).

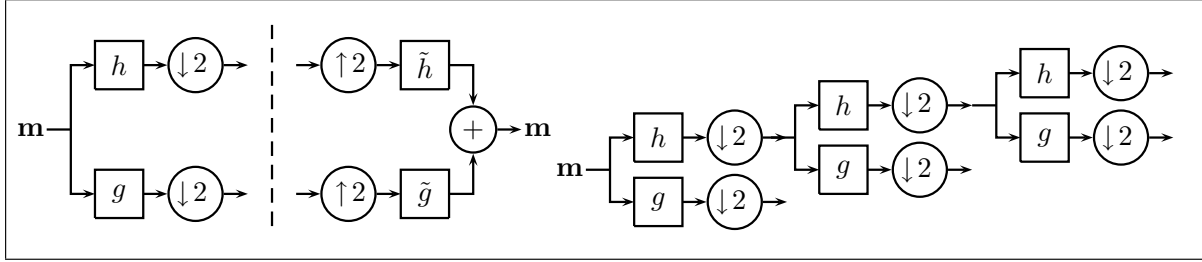| $\nu\,(\mathrm{mHz})$	$C\,(\mathrm{m/s})$	$k\,(10^{-4}\mathrm{m}^{-1})$	$E_0\,(10^{-9}\mathrm{m}^{-2})$	$E_1\,(10^{-9}\mathrm{m}^{-2})$	$E_2\,(10^{-9}\mathrm{m}^{-2})$
10.742	3831.3	0.165 37	−0.079 642	−0.359 72	−0.061 743
15.625	3829.8	0.245 11	−0.126 90	−0.776 64	−0.129 84
20.508	3751.3	0.325 77	−0.176 86	−1.365 6	−0.222 47
30.273	3434.4	0.495 75	−0.368 58	−3.217 7	−0.489 28
40.039	3064.8	0.684 98	−1.078 3	−6.216 1	−0.944 65
50.781	2861.6	0.914 30	−2.788 5	−10.769	−1.815 0
70.313	2872.8	1.344 6	−6.417 5	−22.322	−4.130 0
99.609	2971.5	1.973 3	−11.684	−47.879	−8.884 8

Figure B1: Left: Schematic representation of a perfect-reconstruction filter bank that can be used to decompose or reconstruct a 1D signal x. Right: A standard wavelet tree.

are applied to a given signal $\mathbf{m}$ (i.e. a list of numbers) in the following way: $\mathbf{m}$ is convolved with g and downsampled, $\mathbf{m}$ is convolved with h and downsampled. This results in two signals, each with half the length of the original one. The process can be inverted by upsampling (inserting zeroes) the two resulting sequences and convolving each with two (carefully matched) filters $\tilde{g}$ and $\tilde{h}$ and then adding the two. A traditional way of representing this procedure is shown in the left of Fig. B1. It turns out that there exist *finite* filters that give rise to perfect reconstruction (these use finite convolutions only and lead to compactly supported wavelets); moreover in some very special cases, one can have that finite $\tilde{g}$ and $\tilde{h}$ are the reverse of g and h (corresponding to compactly supported *orthogonal* wavelets). The Haar wavelets have $g = (\frac{1}{2}, -\frac{1}{2})$ and $h = (\frac{1}{2}, \frac{1}{2})$, but there exist longer (perfect reconstruction) finite filters (which give rise to smoother wavelets). The so-called D4 wavelets correspond to $h = (1 + \sqrt{3}, 3 + \sqrt{3}, 3 - \sqrt{3}, 1 - \sqrt{3})/4\sqrt{2}$ and $g = (1 - \sqrt{3}, -3 + \sqrt{3}, 3 + \sqrt{3}, -1 - \sqrt{3})/4\sqrt{2}$.

The 1D discrete wavelet transform is defined by the iteration of the analysis filter bank on the low-pass outcomes (see right side of Fig. B1). In this way, successive levels of detail are stripped of the input signal $\mathbf{m}$ (and stored in wavelet coefficients), leaving a very coarse average (stored in so-called scaling coefficients). This construction is called a wavelet tree. It not only defines the DWT but also provides its practical implementation. When using finite filters, the construction automatically gives rise to a computationally efficient algorithm: as a result of the subsampling each step cost only half as much time as the previous one. The total number of operations then is $kN + kN/2 + kN/4 + kN/8 + \ldots = 2kN$, less than the $\mathcal{O}(N^2)$ for a generic linear transformation.

A standard way of generating wavelets in 2D is to form the direct product of 1D wavelets, i.e. the filters are applied to rows and columns of an image (lo-lo, lo-hi, hi-lo and hi-hi). This, however, has the marked disadvantage of poor directional sensitivity. In this study, to obtain better directional sensitivity,
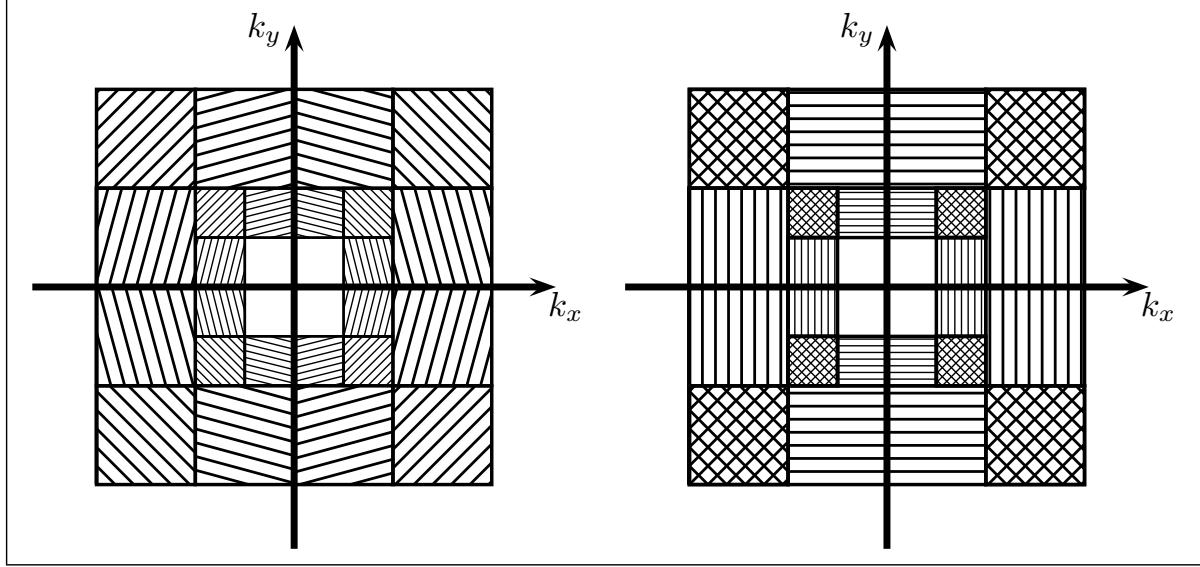
Figure B2: Partitioning of the 2D Fourier domain (k_x, k_y) by the supports of the Fourier transform of wavelet functions. Only two wavelet scales are shown with the finest one on the outside. In practice the supports have smooth (overlapping) tapers. Left: With the complex 2D wavelets used in this paper, all squares come in pairs giving rise to six dominant directions (as indicated by the hatch patterns). Right: The standard (direct product) 2D construction only has horizontal and vertical sensitivity; the 'corner' (hi-hi) squares encode both $45°$ and $-45°$ at the same time.

we use the complex 2D wavelets developed by [Kingsbury(2002)]. These are constructed also by direct product but from *two* simultaneous wavelet trees (see [Kingsbury(1999)] for a diagram of such a dual tree). The qualitative difference between these two constructions is best seen in the Fourier domain. Fig. B2 shows a schematic representation of the supports of the Fourier transforms of the wavelet functions, both for the usual 2D wavelet construction and for the 2D complex wavelets. The two are fundamentally different: whereas the usual separable 2D wavelet construction gives rise to a horizontal, a vertical and one (!) diagonal part at each scale, the complex 2D construction has six different inherent directions per scale. A careful choice of the different filters also leads to an (almost) tight frame (i.e. the inverse wavelet transform almost coincides with the transpose).

The price to pay for these benefits is the redundancy. In 2D the complex wavelets generate four times as many coefficients as there are pixels in the original image (two trees and real and imaginary parts of the output). E.g. the 64×64 spatial-domain images we use in Section 3 give rise to $16\,320 = 2 \times 6 \times (32^2 + 16^2 + 8^2 + 4^2)$ wavelet coefficients and $64 = 2 \times 2 \times 4^2$ scaling coefficients (see e.g. Fig. 9). Together this is $16\,384$ which equals 4×64^2.